protoDUNE DP online computing

Elisabetta Pennacchio, IT-protoDUNE coordination meeting, 08/03/2017

These slides are a summary of what has already been presented in different meetings, and are based on:

- the DAQ system architecture
- the experience already acquired with the storage and processing farm setup and commissioning for the 3x1x1.

These slides summarize the work of several people in particular of:

- Denis Pugnere (IPNL) who has performed extensive studies on the design of the local storage system and network architecture

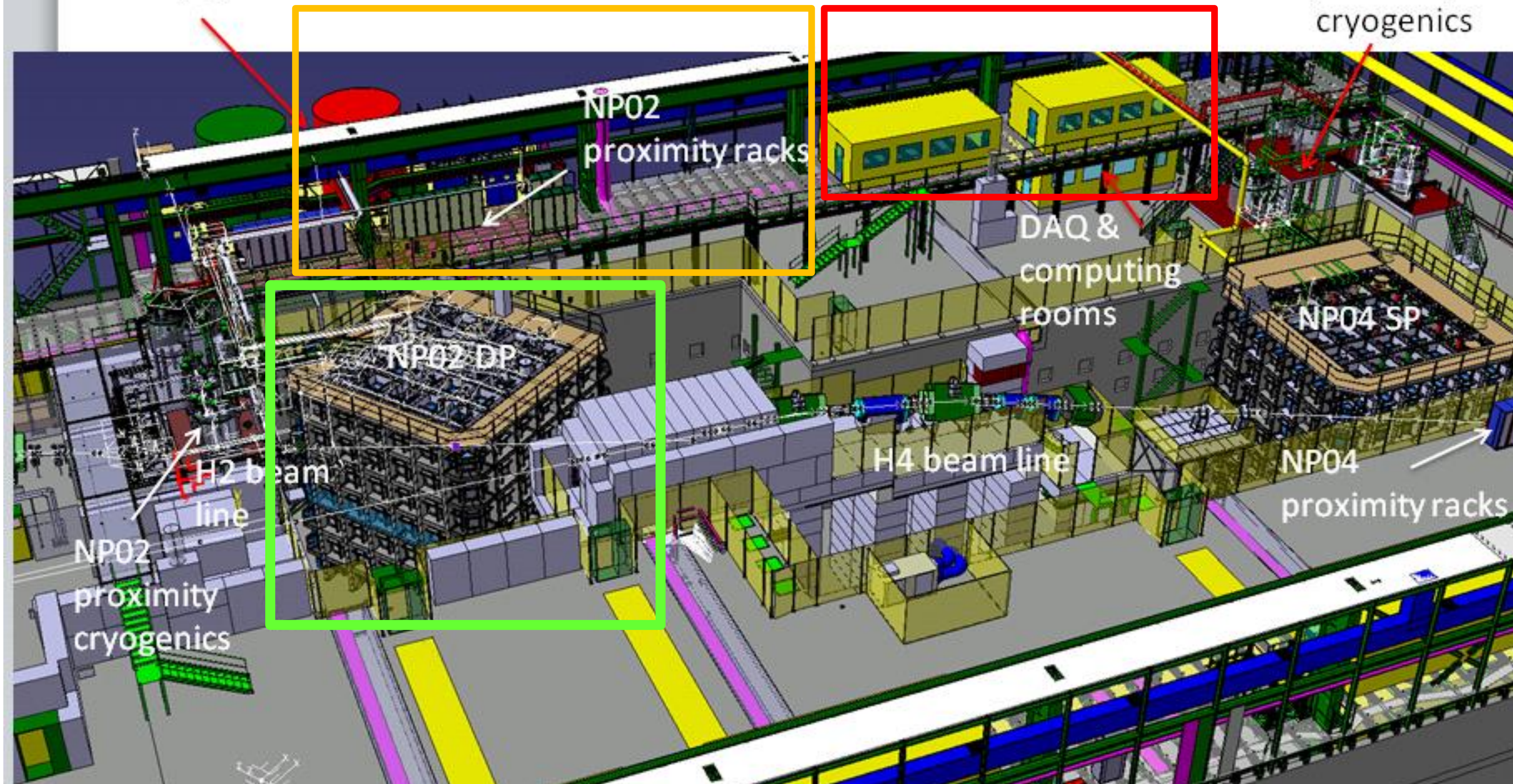
- Thierry Viant (ETHZ) has setup all the network infrastructure and the slow control database

- The Lyon group which has the responsibility of the DAQ (for both 3x1x1 and 6x6x6),

- The Kiev group which has developed the slow control database WEB interface

- Introduction
- DAQ architecture
- 3x1x1 implementation
- Trigger timing distribution white rabbit network
- Beam instrumentation/ProtoDUNE-DP synchronization and interface
- Online storage and processing

EHN1-1 Layout



CATIA, integration model

Double phase liquid argon TPC
6x6x6 m3 active volume

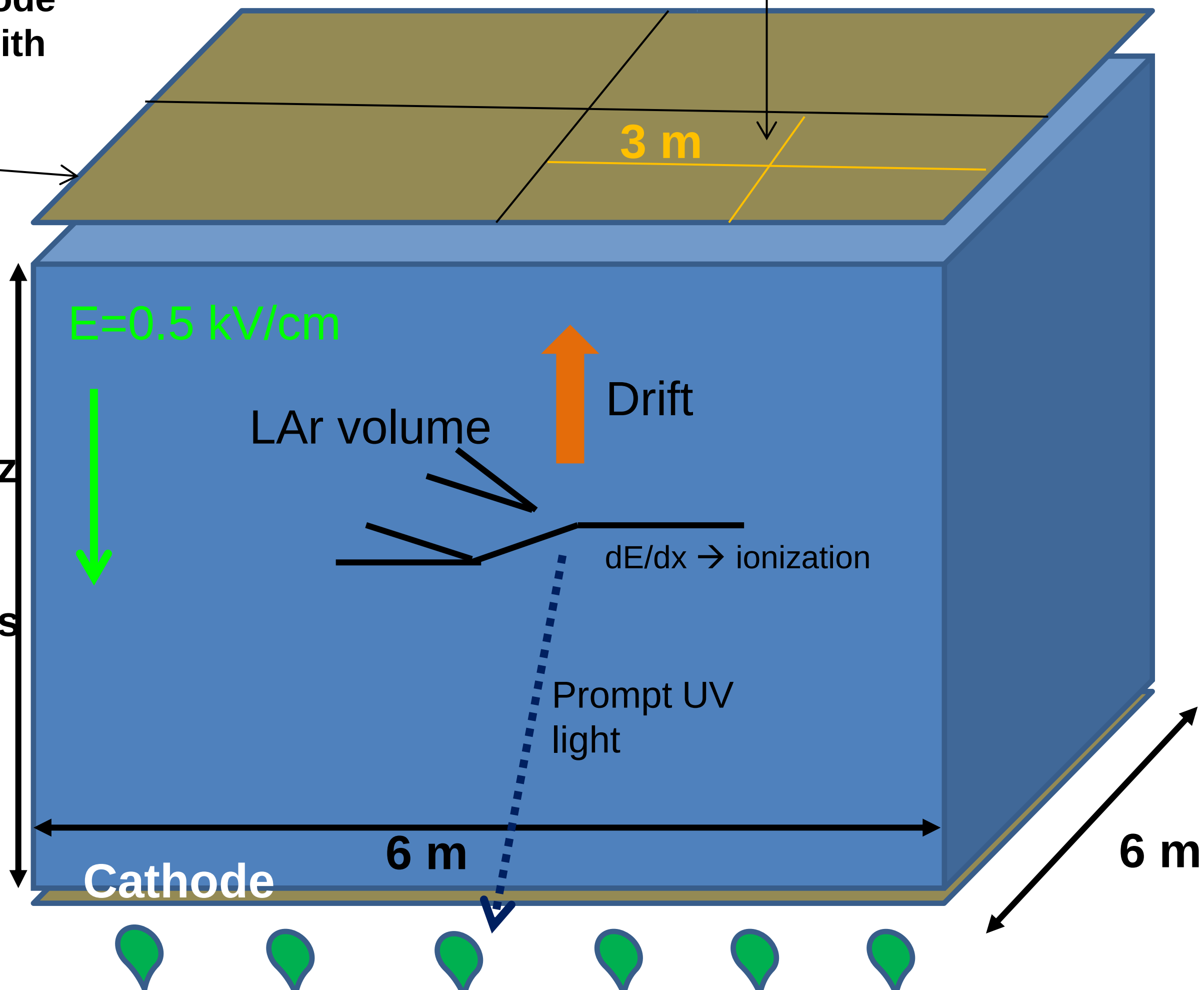
→ **Event size: drift window of 7680 channels x 10000 samples = 146.8 MB**

X and Y charge collection strips
3.125 mm pitch, 3 m long
→ 7680 readout channels

Segmented anode In gas phase with double-phase amplification

Drift coordinate
6 m = 4 ms
sampling 2.5 MHz (400 ns), 12 bits

→ 10000 samples per drift window



WA105 Accessible cold front-end electronics and uTCA DAQ system 7680 ch

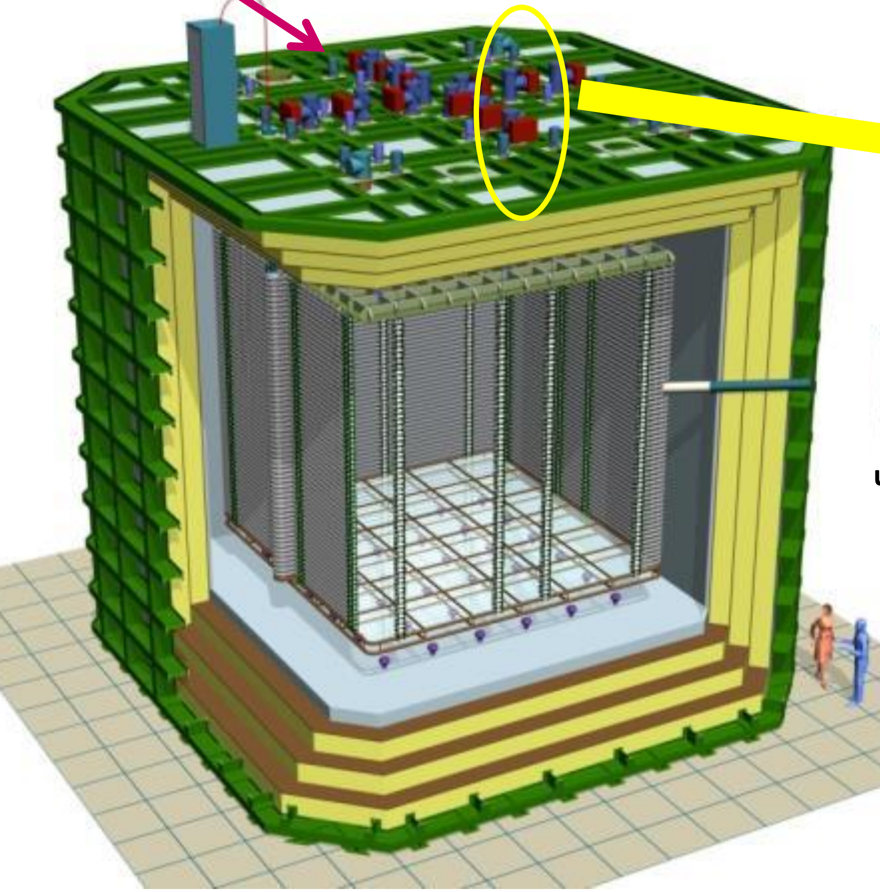
Full accessibility provided by the double-phase charge readout at the top of the detector

- **Digital electronics at warm on the tank deck:**
  - Architecture based on uTCA standard
  - 1 crate/signal chimney, 640 channels/crate

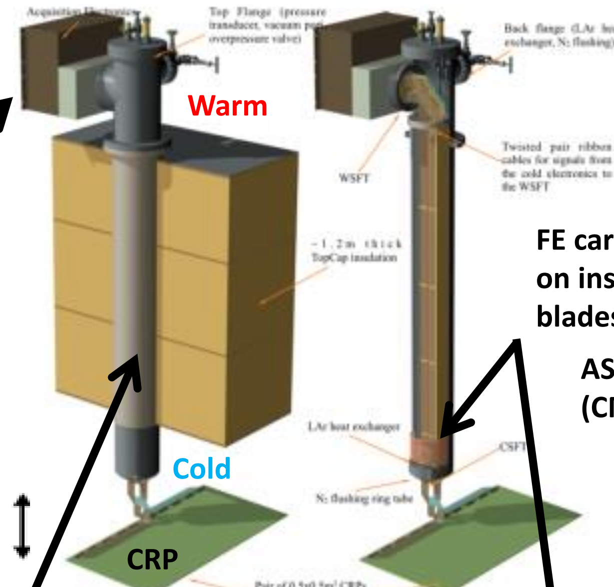
→ 12 uTCA crates, 10 AMC cards/crate, 64 ch/card

- **Cryogenic ASIC amplifiers (CMOS 0.35um) 16ch externally accessible:**
  - Working at 110K at the bottom of the signal chimneys
  - Cards fixed to a plug accessible from outside

→Short cables capacitance, low noise at low T

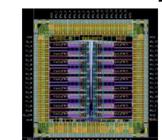


uTCA crate

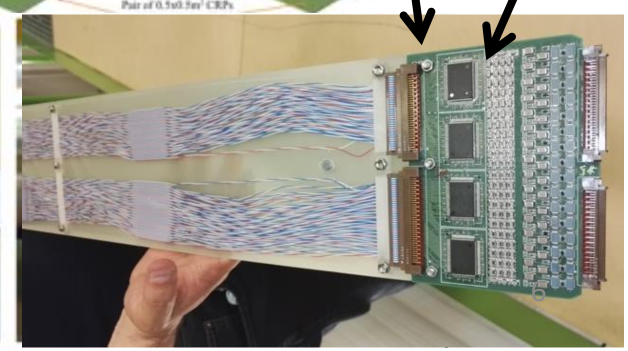


FE cards mounted on insertion blades

ASICs 16 ch. (CMOS 0.35 um)



Signal chimney



DAQ architecture

- **Dual phase ProtoDUNE detector characteristics:**
- Two views with 3.125 mm pitch → 7680 channels
- Long drift 4 ms → 10000 samples at 2.5 MHz
- High S/N~100

- **All electronics at warm, accessible**

→ **Costs minimization**, massive use of **commercial** large bandwidth standards in telecommunication industry, **uTCA**, Ethernet networks, massive computing

→ Easy to follow technological evolution, benefit of costs reduction and increase of performance in the long term perspective

- **Non-zero suppressed data flow** handled up to computing farm back-end which is taking care of final part of event building, data filtering, online processing for data quality, purity, gain analysis, local buffer of data and data formatting for transfer to EOS storage in files of a few GBs

- Signal lossless compression benefits by high S/N ratio, developed an optimized version of Huffman code reducing data volume by at least a factor 10

- Timing and trigger distribution scheme based on White Rabbit (became commercial hardware too); thought since the beginning for **a beam application** (handles beam window signals, beam trigger counters, external trigger counters for cosmics) → Components of the timing chain purchased and uTCA slave card for signal distributions on crates backplane developed

- FE based on microTCA standard and 10 Gbit/s ethernet

→ 120 uTCA digitization boards went under production for the 6x6X6 since 2015, 20 card already installed on 3x1x1

- Light readout fully integrated in charge readout uTCA scheme and with different operation modes in-spill out-spill

- Back-end actually based on two commercial Bittware cards with x8 10 Gbit links with high computing power and event building capabilities. Each card performs event building for ½ of the detector charge readout, one card deals with light signals too

- Online storage and computing facility is an important part of the system, a possible implementation has been designed and costed with DELL, it has been implemented on a smaller scale for the 3x1x1 (September 2016)

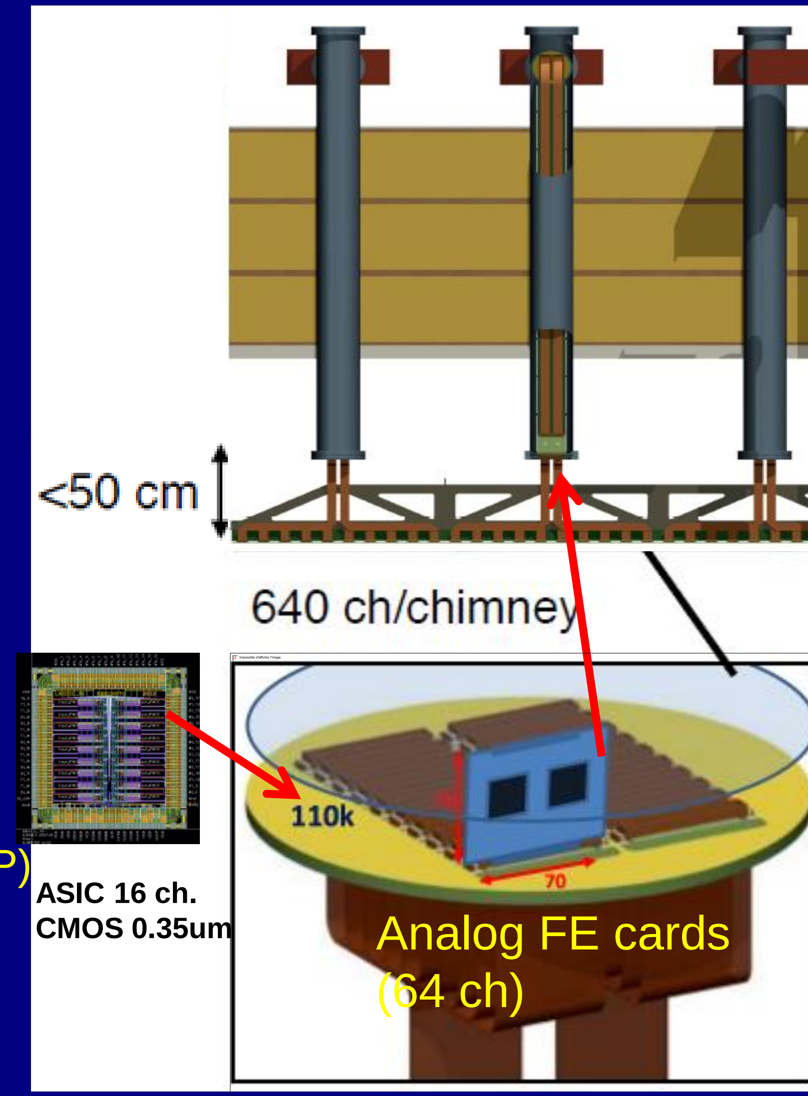
- 20 Gbit/s data link foreseen for data transfer to computing division

Cost effective and fully accessible cold front-end electronics and DAQ

Ongoing R&D since 2006→ in production for 6x6x6 (7680 readout channels)

ASIC (CMOS 0.35 um) 16 ch. amplifiers working at ~110 K to profit from minimal noise conditions:

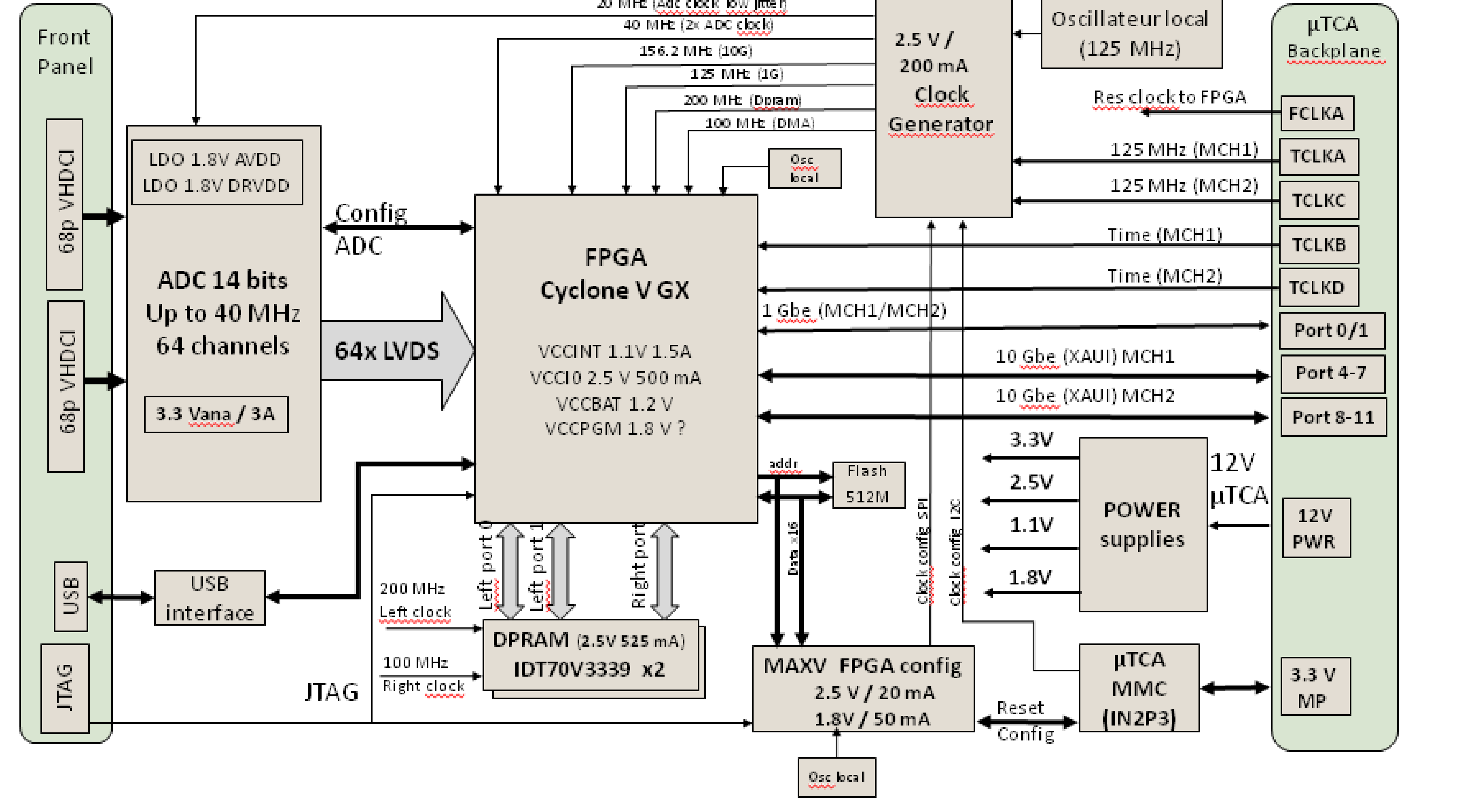
- FE electronics inside chimneys, cards fixed to a plug accessible from outside
- Distance cards-CRP<50 cm
- Dynamic range 40 mips, (1200 fC) (LEM gain =20)
- 1300 e- ENC @250 pF, <100 keV sensitivity
- Single and double-slope versions
- Power consumption <18 mW/ch
- **Produced at the end of 2015 in 700 units (entire 6x6x6)**
- **1280 channels installed on 3x1x1**



DAQ in warm zone on the tank deck:

- Architecture based on uTCA standard
- Local processors replaced by virtual processors emulated in low cost FPGAs (NIOS)
- Integration of the time distribution chain (improved PTP)
- Bittware S5-PCIe-HQ 10 Gbe backend with OPENCL and high computing power in FPGAs
- **Production of uTCA cards started at the end of 2015, pre-batch already deployed on 3x1x1**

→ Large scalability (150k channels for 10kton) at low costs

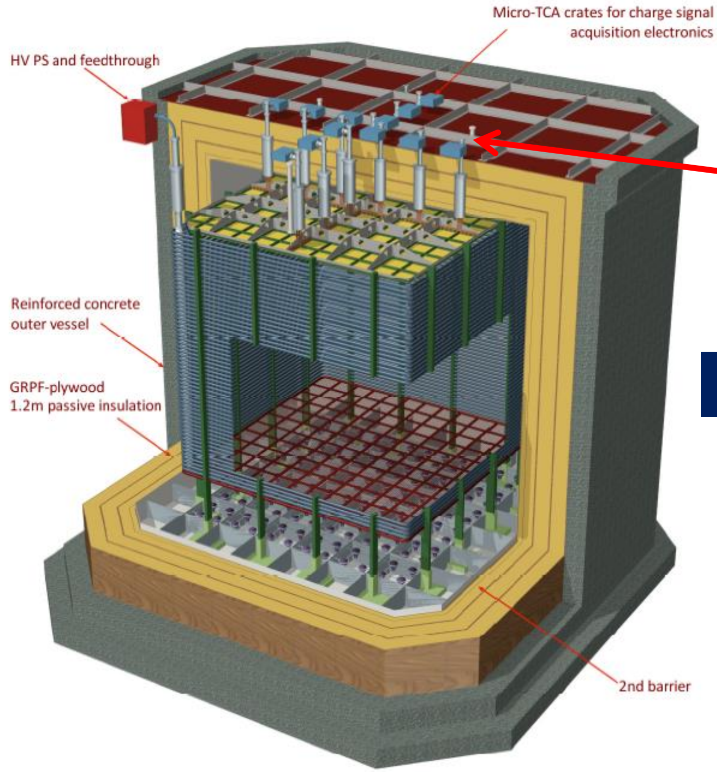


Production of uTCA AMC cards for the 6x6x6 started at the end of 2015, first batch deployed on 3x1x1

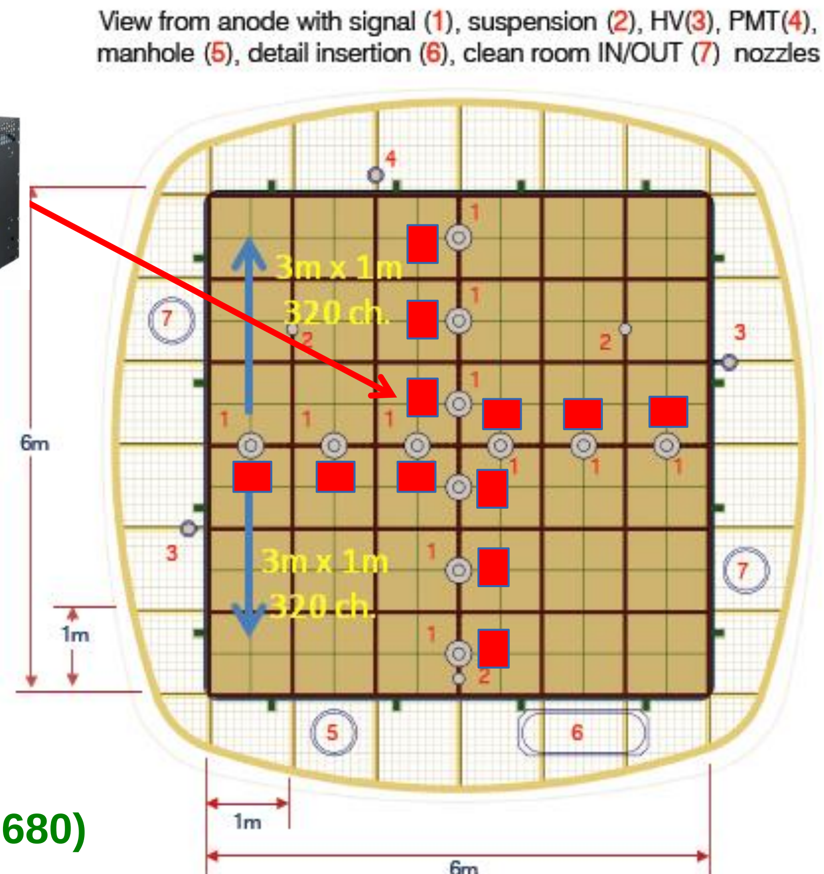
First 64 ch AMC digitization card delivered at the end of July 2016

(2.5-25 MHz, 12 bits, 2V, ADC AD5297, 10 GbE output on uTCA backplane)

uTCA charge readout architecture



uTCA crates



Readout in groups of 640 channels/chimney

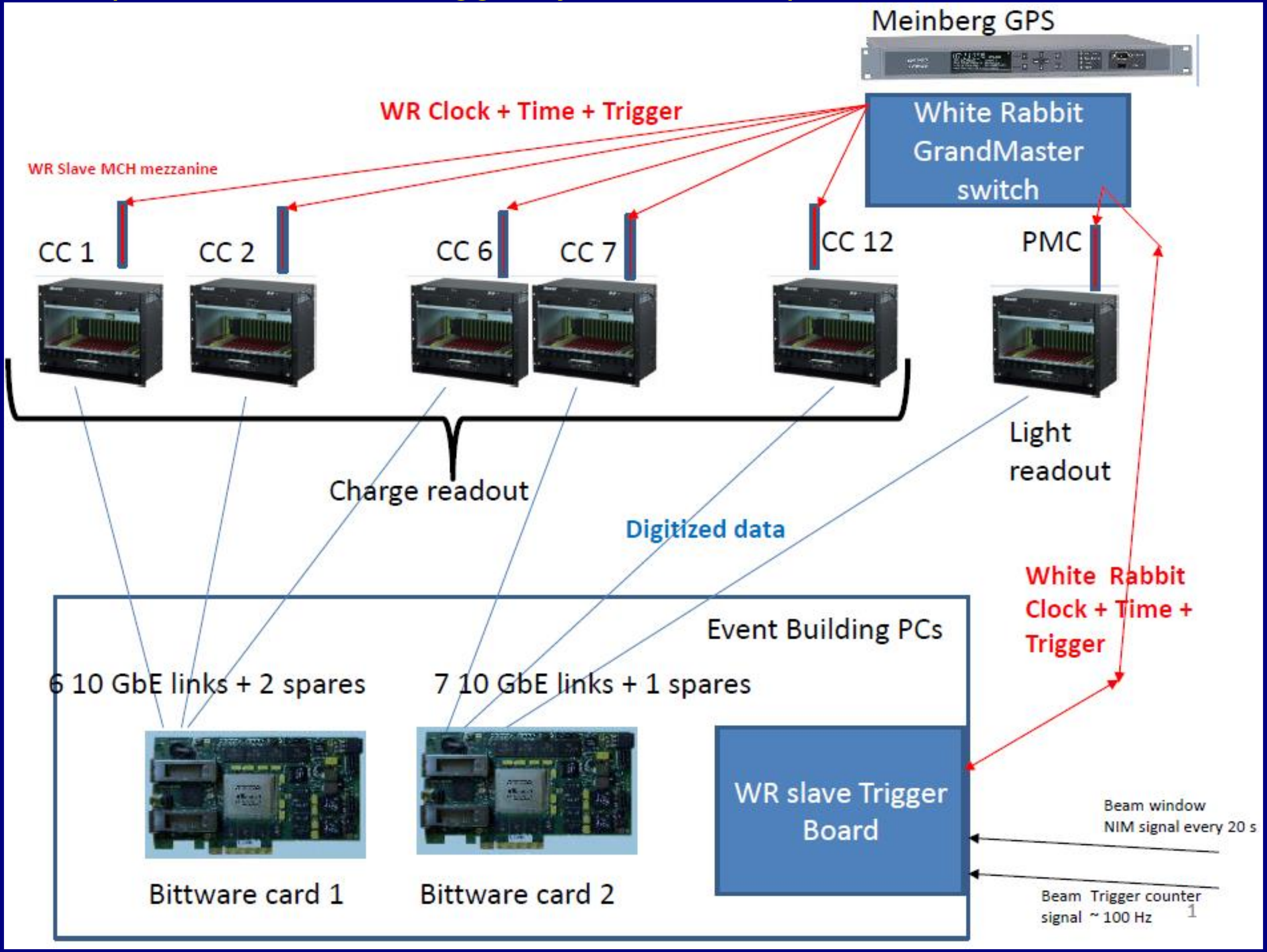
top view

- 12 chimneys, 640 channels/chimney (640*12=7680)
- 1 uTCA crate/chimney, 10 Gb/s link
- 10 AMC digitization boards per uTCA crate, 64 readout channels per AMC board

→ 12 uTCA crates for charge readout + 1 uTCA crate for light readout

Global uTCA DAQ architecture

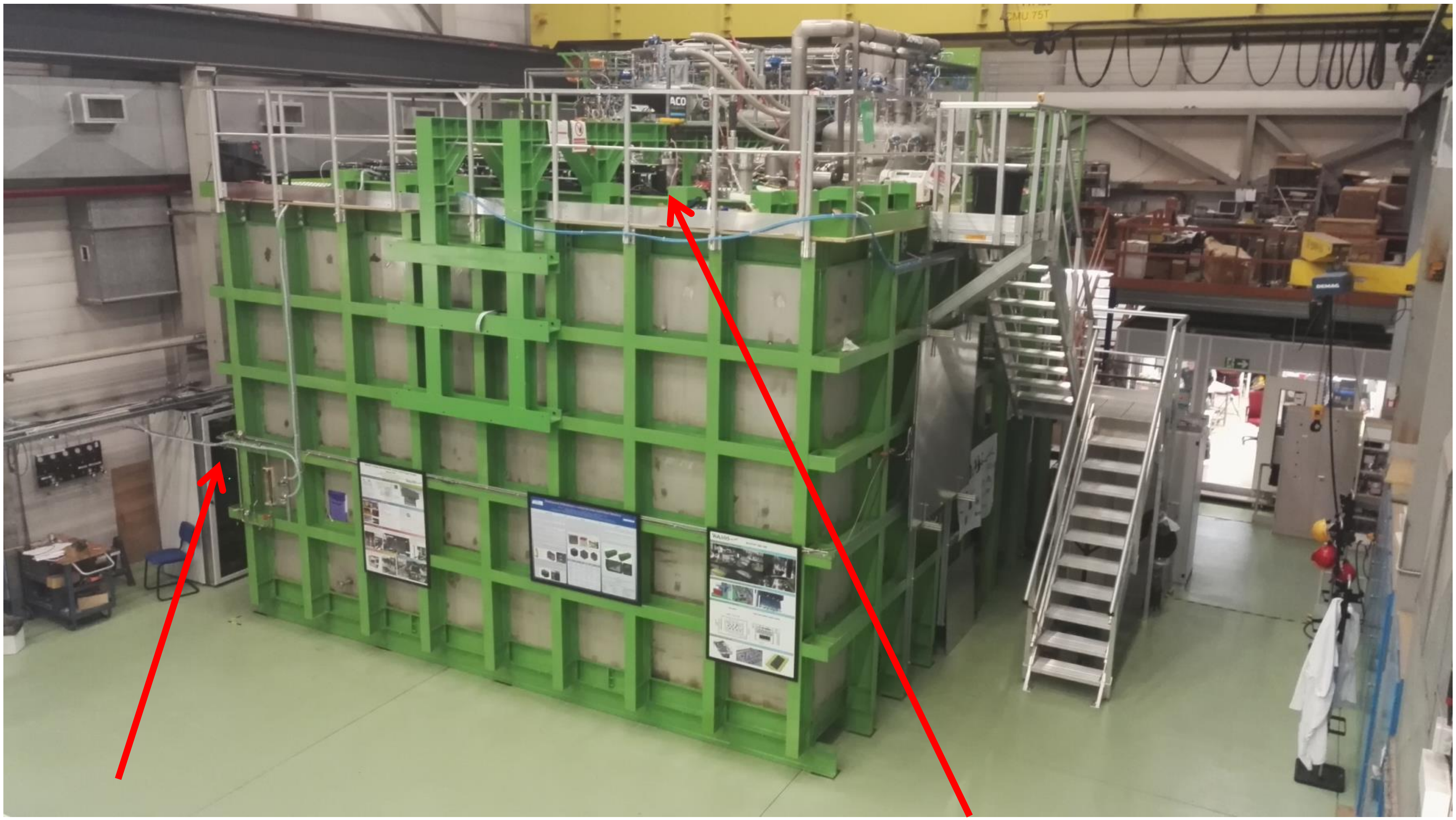
integrated with « White Rabbit » (WR) Time and Trigger distribution network
+ White Rabbit slaves nodes in uTCA crates +
WR system (time source, GM, trigger system, slaves)



3x1x1 implementation

6x6x6: 12 uTCA crates (120 AMCs, 7680 readout channels)

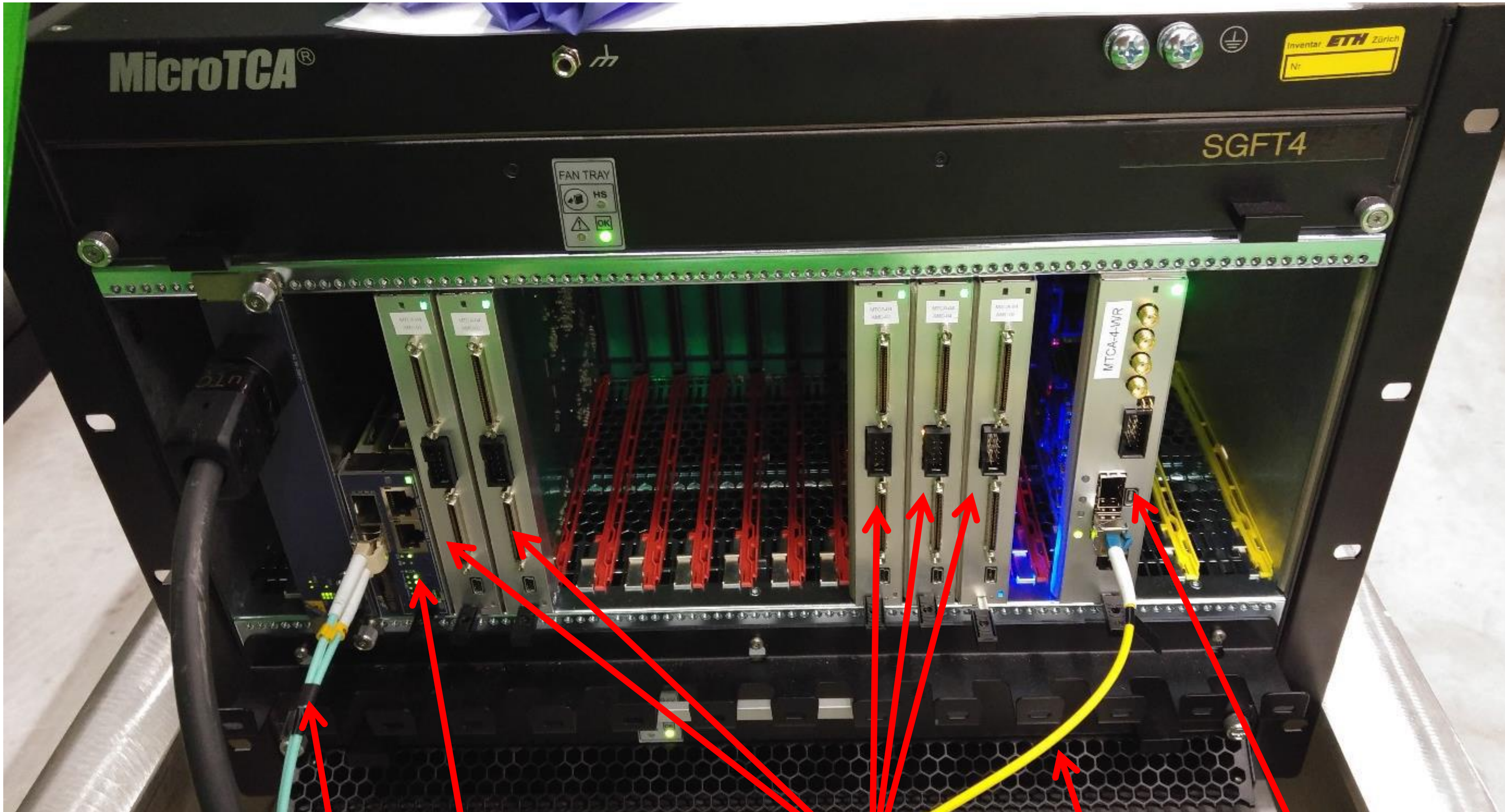
→ 3x1x1: 4 uTCA crates (20 AMCs, 1280 readout channels)
+ Slow Control

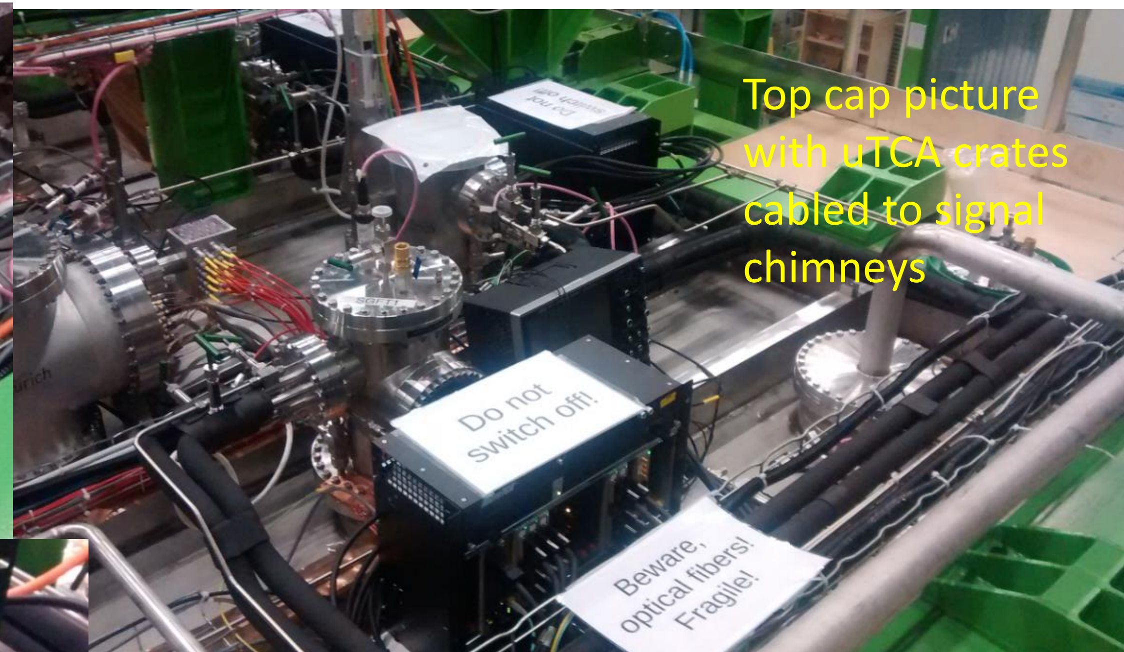


Event builder, network, GPS/White Rabbit GM, WR Trigger PC

Signal Chimneys and uTCA crates

How a crates was looking like before VHDCI signals cabling to the warm flange

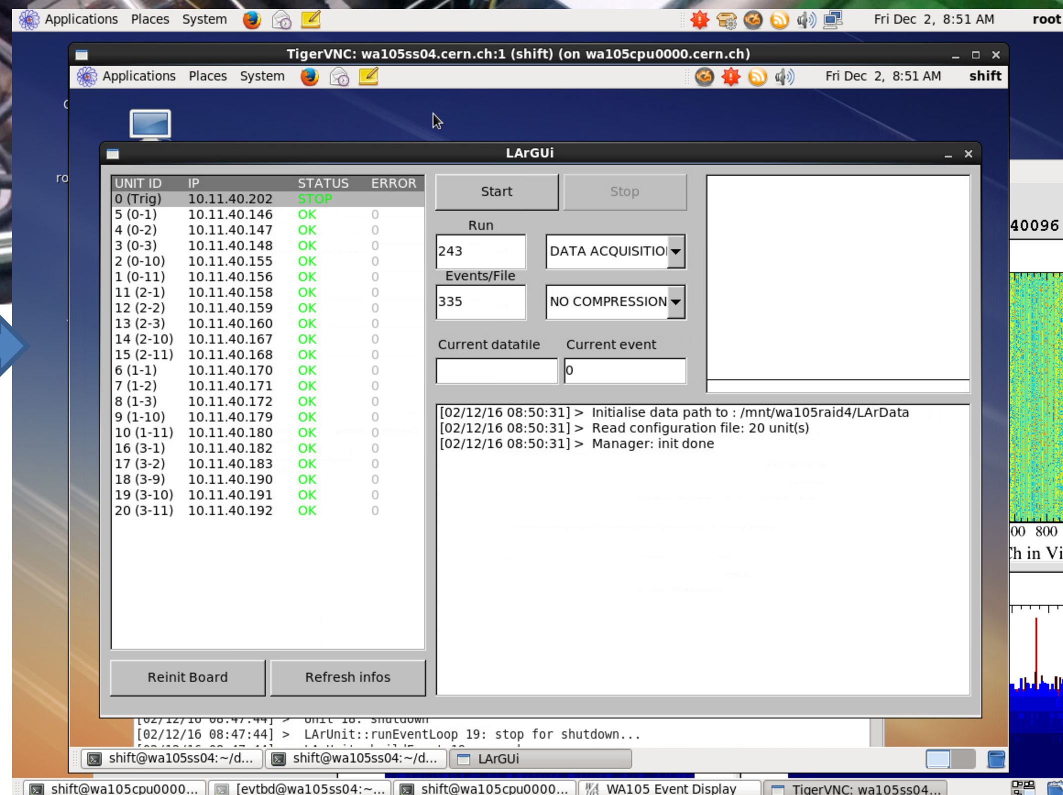




Run control with 20 AMCs

Automatic data processing on online storage/processing farm for purity and gain analysis + data transfer on EOS

Stable system, noise conditions at warm 1.5-2.4 ADC counts RMS



Trigger timing distribution white rabbit network

- White Rabbit developed for the synchronization of CERN accelerators chain offers sub-ns synchronization over ~10 km distance, based on PTP + synchronous Ethernet scheme previously developed in 2008 (http://arxiv.org/abs/0906.2325)

- White Rabbit chains can be now set up with commercial components:
  - Network based on Grand Master switch
  - Time tagging cards for external triggers
  - Slave nodes in piggy back configuration to interface to uTCA

- Transmission of synchronization and trigger data over the WR network + clock

- Slave uTCA nodes propagate clock + sync + trigger signals on the uTCA backplane, so that the FE digitization cards are aligned in their sampling, can know the absolute time and can compare it with the one of transmitted triggers

- FE knows spill time and off spill time and can set up different operation modes

- Trigger timestamps may be created by beam counters, cosmic counters, light readout system in uTCA

- The White Rabbit network provides the distribution of a common time base to align all the elements of the DAQ system: the 400 ns sampling on the uTCA digitization boards of the charge readout and the light readout digitization boards

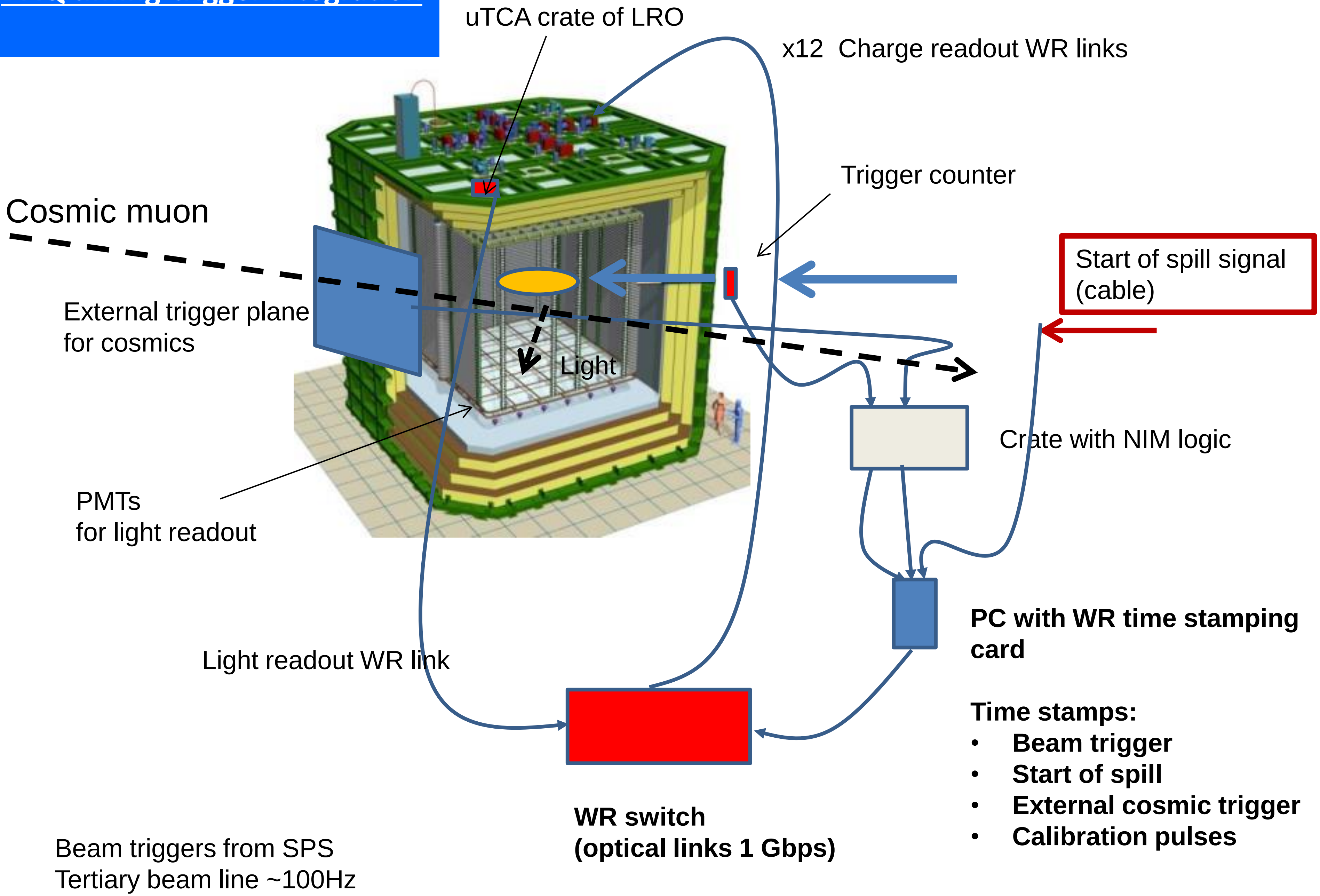
- The White Rabbit network is also used to distribute triggers (timestamps of trigger signals in this common time base) to the elements of the DAQ chain. The uTCA digitization boards have a large circular memory buffer (larger than the drift time) and associate to the drift window the samples starting from the time stamp of the triggers

- Triggers are created and injected in the network by a WR timestamping board in the trigger PC. This looks at 3 input logic signals (connections via LEMO cables):

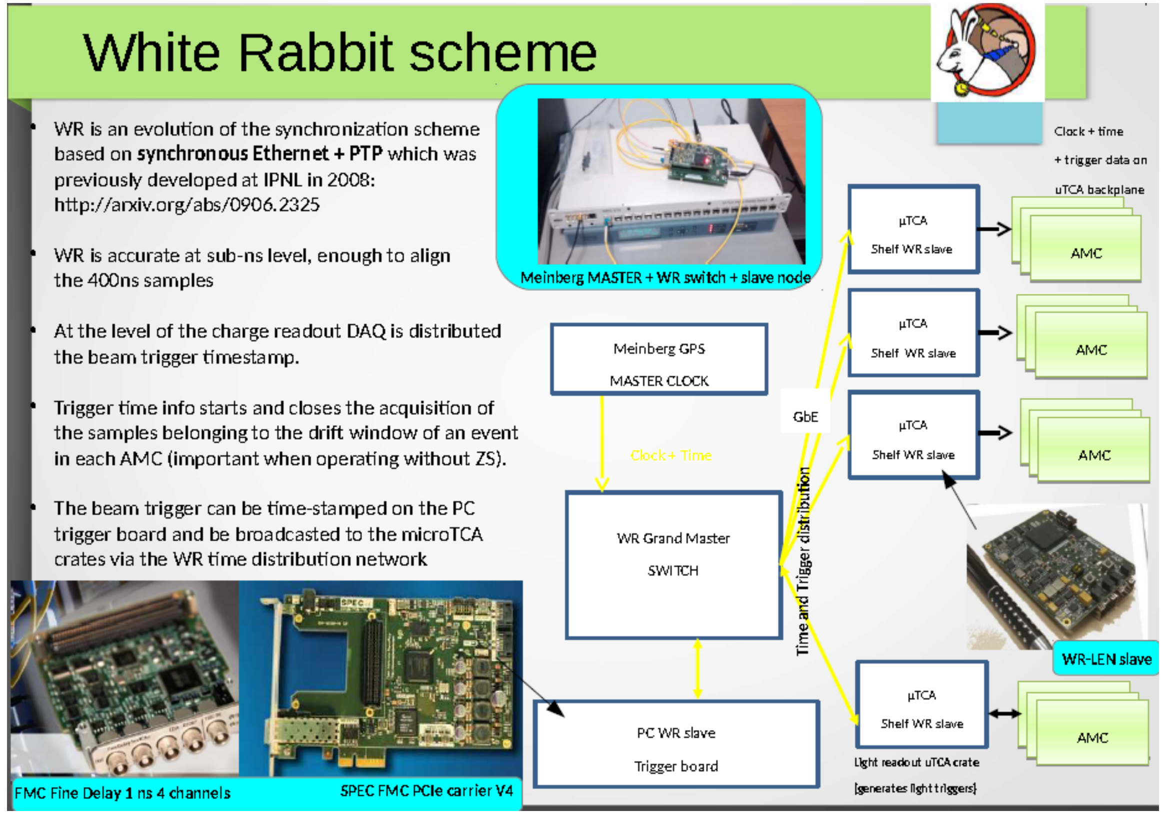
a) The start of spill signal (the FE needs to know if it is taking data during or out of the spill in order to deal with different triggers and set different sampling modes of the LRO)
b) Beam trigger (from the scintillators along the beam line)
c) Cosmic ray taggers

The LRO triggers should be directly injected in the WR network

DAQ timing-trigger integration



White Rabbit trigger time-stamping PC (SPEC + FMC-DIO)
White Rabbit Grand-Master
GPS unit



White Rabbit uTCA slave node based on WRLEN developed and produced for entire 6x6x6

Other components of the chain (GPS receiver, WR grandmaster, SPEC+ FMC-DIO + 13 WRLEN) available commercially

Beam instrumentation/ProtoDUNE-DP synchronization and interface

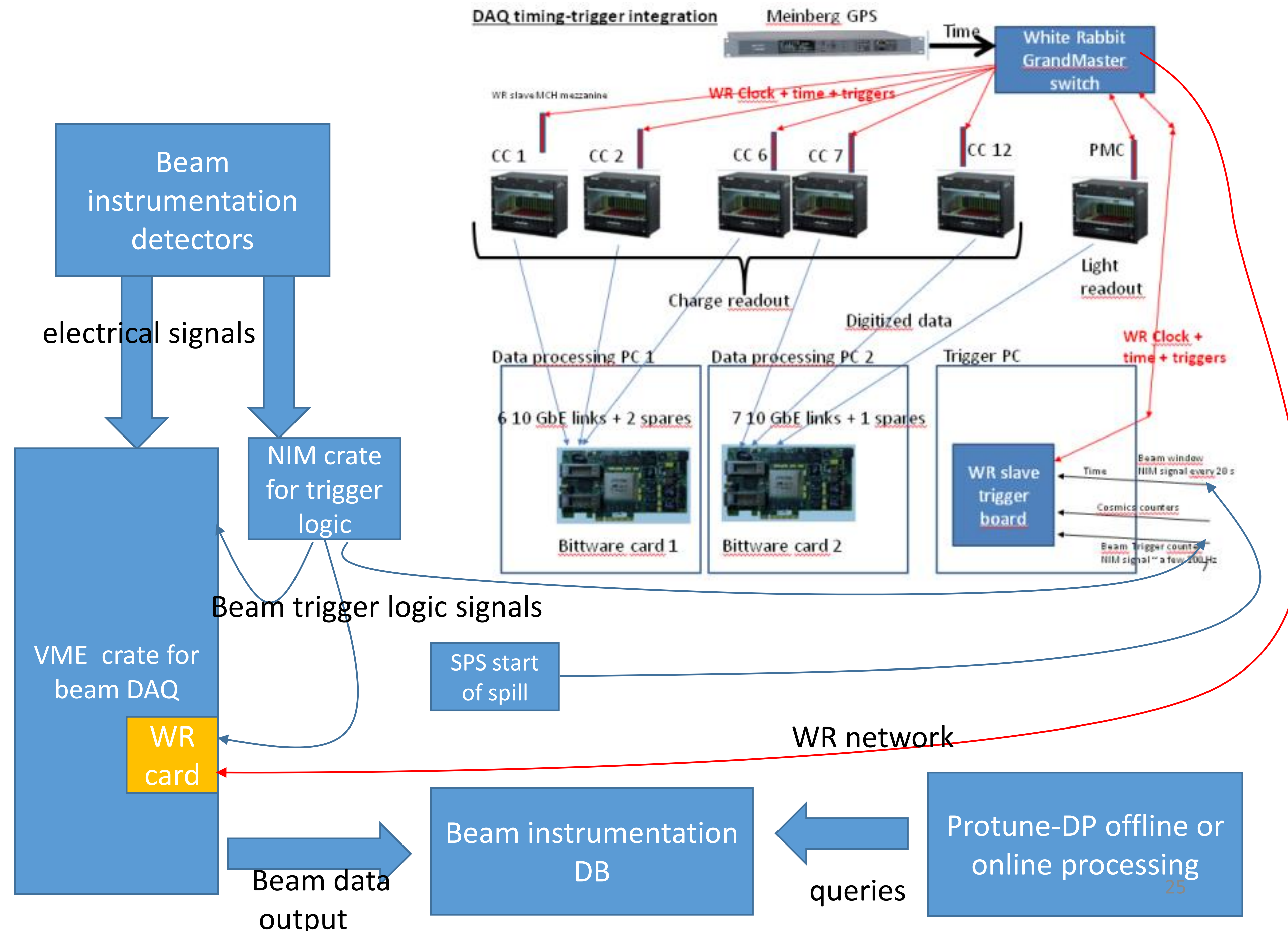
The DAQ system of the beam instrumentation will also be triggered by the NIM signal from trigger scintillators and will take data for the different ADC/TDC/IO registers (in VME) used to read the beam instrumentation. The NIM crate which defines the coincidence for the NIM trigger will have a fan-out in order to distribute the trigger signal to both the local beam DAQ PC, the beam DAQ crate and to the ProtoDUNE-DP trigger PC

The DAQ VME system will have also a WR time tagging card installed which looks at the beam trigger. This WR card is connected via an optical fiber to the WR grand master of WA105 so that it is aligned on the common time base and it is read out with the beam DAQ. Once the beam DAQ system reads the event from the beam instrumentation it will also read the timestamp of the beam trigger and associate to the event structure

The beam DAQ will write the beam trigger data from the local beam instrumentation DAQ which includes the time tagging WR card on the beam instrumentation database

The online computing farm (or any offline process) can access the beam instrumentation database in order to fish the beam instrumentation data related to a given timestamp for a ProtoDUNE-DP trigger

From the beam line to protoDUNE-DP two cables one for the beam trigger signal and one for the start of spill signal. From WA105 to the Beam DAQ we will deploy an optical fiber with the WR network connection → General scheme in the next page:

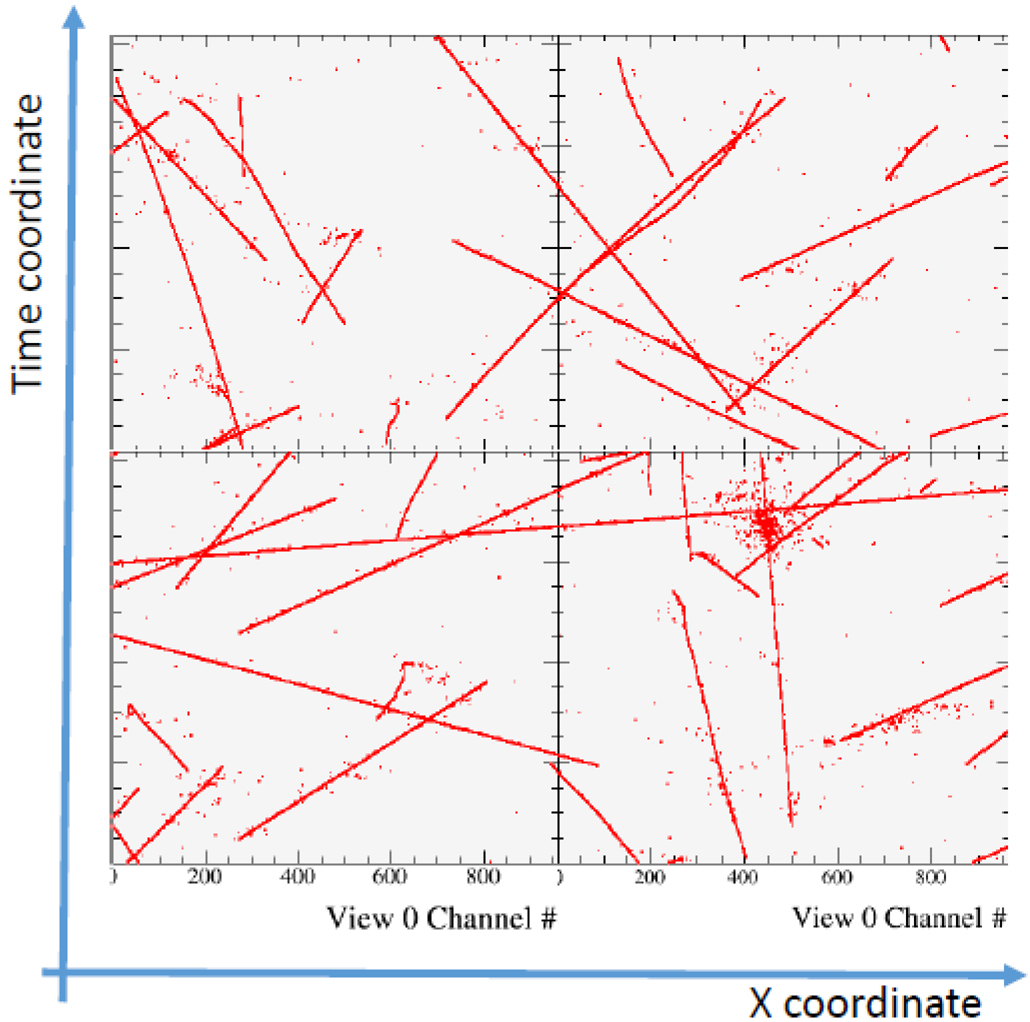

DAQ timing-trigger integration
Meinberg GPS
Time
White Rabbit GrandMaster switch
WR slave MCH mezzanine
WR Clock + time + triggers
CC 1
CC 2
CC 6
CC 7
CC 12
PMC
Light readout
Charge readout
Digitized data
WR Clock + time + triggers
Data processing PC 1
Data processing PC 2
Trigger PC
6 10 GbE links + 2 spares
7 10 GbE links + 1 spares
Bittware card 1
Bittware card 2
WR slave trigger board
Time
Beam window
NIM signal every 20 s
Cosmics counters
Beam Trigger counter
NIM signal ~ a few 100 Hz
Beam instrumentation detectors
electrical signals
NIM crate for trigger logic
Beam trigger logic signals
VME crate for beam DAQ
WR card
SPS start of spill
WR network
Beam data output
Beam instrumentation DB
queries
Protune-DP offline or online processing

Online storage and processing

Typical event signature for ground surface Liquid Ar TPC operation

For each beam trigger we can have on average 70 cosmics overlapped on the drift window after the trigger (these cosmics may have interacted with the detector in the 4 ms before the trigger and in the 4 ms after the trigger → chopped tracks, "belt conveyor" effect

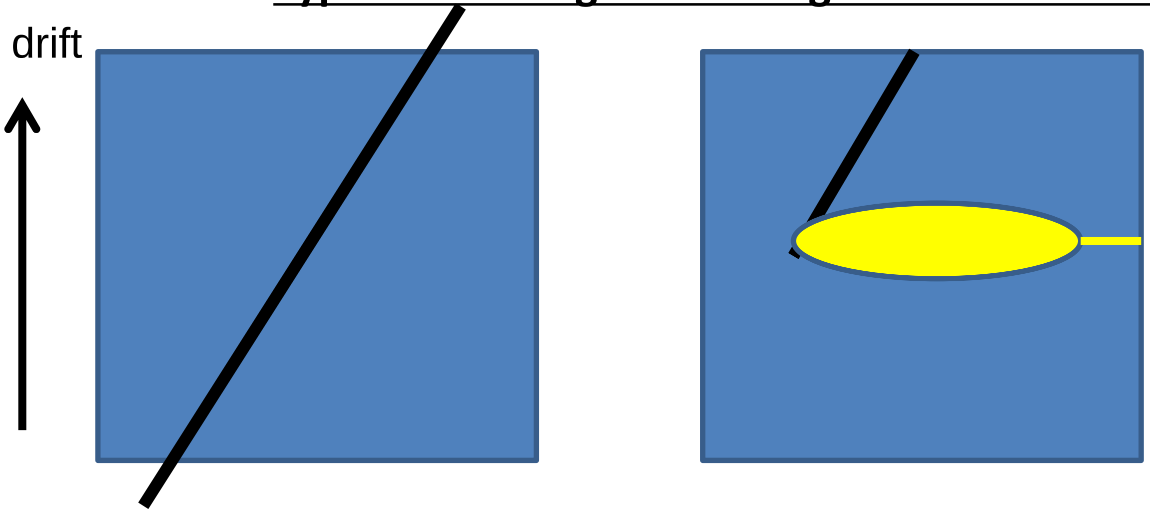
In-spill cosmics in charge data



Example of cosmics only event (in one of the views)

- Red points are reconstructed hits
- TPC is readout in 4 3x3m^2 modules
- After track reconstruction:
  - Attempt to correlate found tracks with light data
  - Remove CR background from beam event
  - Select a subsample of long tracks for calibration purposes

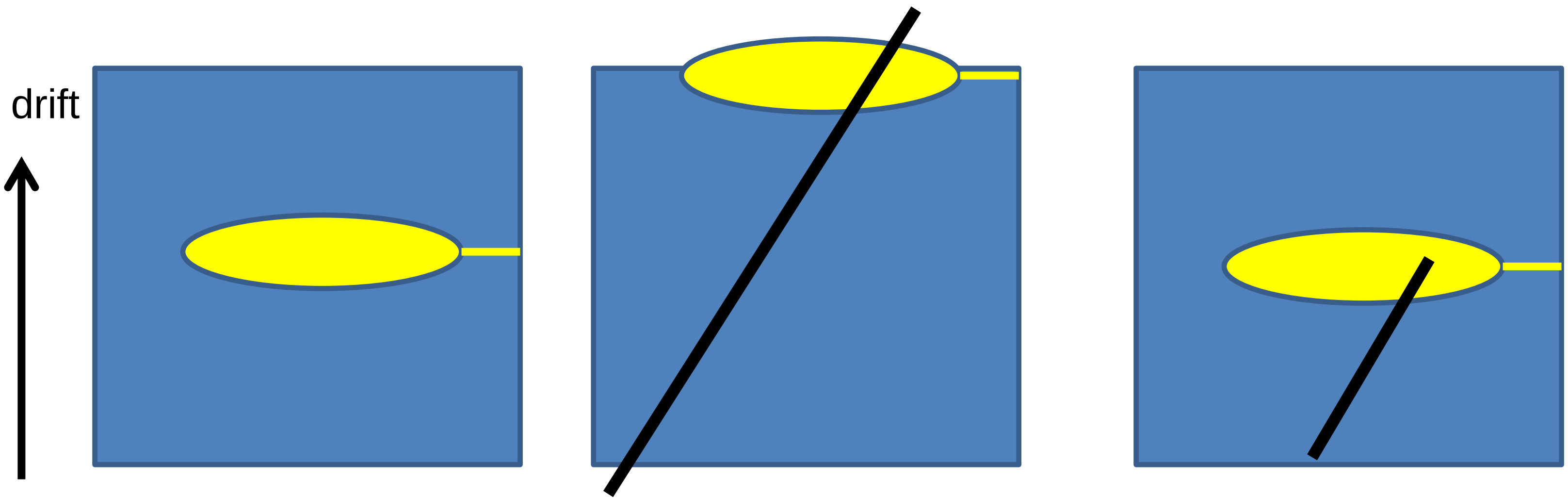
Typical event signature for ground surface Liquid Ar TPC operation



t=beam trigger - 2 ms

t=beam trigger → reconstructed event

The « belt conveyor » effect
+- 4 ms around the beam
trigger time



t=beam trigger

t=beam trigger + 2 ms

reconstructed event

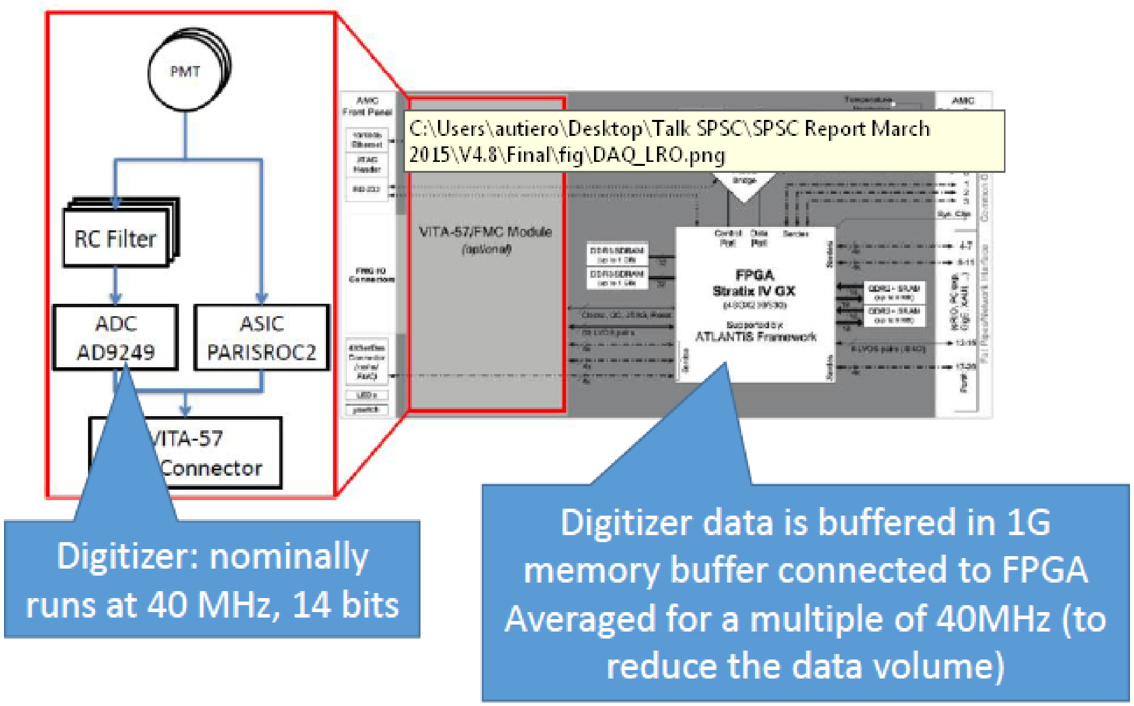
→ During spills it is needed a continuous digitization of the light in the +-4 ms around the trigger time (the light signal is instantaneous and keeps memory of the real arrival time of the cosmics)

→ Sampling can be coarse up to 400 ns just to correlate to charge readout

Light readout electronics

Two modes of acquisition:

- External beam trigger to acquire ±4ms around the spill
- Internal trigger from PARISROC2 ASIC to acquire short time segments



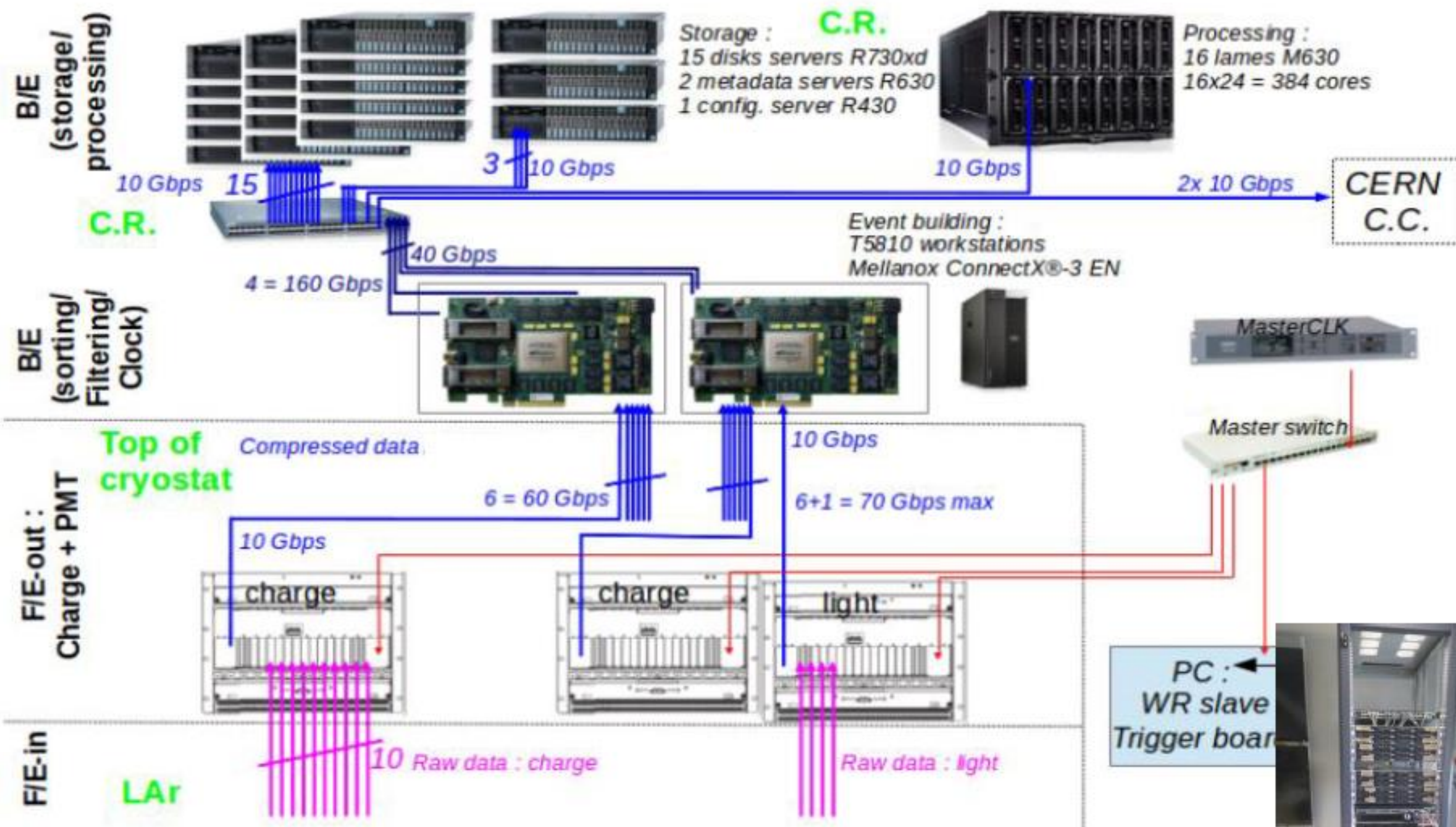
→ Sum 16 samples at 40MHz to get an effective 2.5 MHz sampling like for the charge readout

The LRO card has to know spill/out of spill

Out of spill it can define self-triggering light triggers when "n" PMTs are over a certain threshold and transmit its time-stamp over the WR

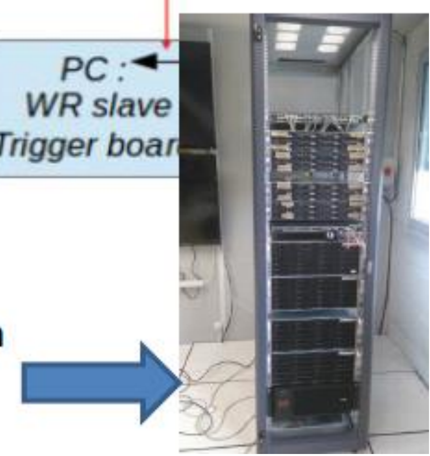
DUNE meeting September 2016: **Online processing and storage facility of 6x6x6**

Online processing and storage facility: internal bandwidth 20 GB/s, 1 PB storage, 384 cores: key element for online analysis (removal of cosmics, purity, gain, events filtering)



C.R. stands for Counting Room

Smaller scale test system implemented for the operation of the 3x1x1



▪Design of online storage/processing DAQ back-end farm completed in 2016 (1PB, 300 cores, 20Gb/s data flow),

DELL-based solution : configuration

storage servers :
* 15 R730XD (storage servers) including :
* 16 disks 6To
* 32Go RAM
* 2 disks system RAID 1, 300 Go 10k
* 1 network card Intel X540 double port 10 GB
* 4 years extended guarantee (D+1 intervention)
* 2 processors Intel Xeon E5-2609 v3
* raid H730P
* Rails with management arm
* double power supply

metadata servers (MDS) :
* 2 R630 (metadata servers), including :
* 2 disks 200 Go SSD SAS Mix Use MLC 12Gb/s
* 2 processors Intel Xeon E5-2630 v3
* 32Go DDR4
* RAID H730p
* network : Intel X540 2 ports 10 Gb
* 4 years extended guarantee (D+1 intervention)
* Rails with management arm
* double power supply

configuration server :
* 1 R430 (configuration server)
* 1 processor E5-2603 v3
* RAID H730
* 2 hard disks 500 Go Nearline SAS 6 Gbps 7,2k
* 16 Go DDR4
* Rails with management arm
* double power supply

Offline computing farm: 16*24 = 384 cores
* 1 blade center PowerEdge M1000e with 16 lames M630, each including :
  * 128Go DDR4
  * 2 processors Intel Xeon E5-2670 v3
  * 4 years extended guarantee (D+1 intervention)
  * 2 hard disks 500 Go SATA 7200 Tpm
  * netwok Intel X540 10 Gb

Switch Force10, S4820T (see next slide) :
* 48 x 10GbaseT ports
* 4 x 40G QSFP+ ports
* 1 x AC PSU
* 2 fans

▪Prototype already installed and operative for 3x1x1
Tests to finalise the architecture of final farm

▪ 5 Storage servers 240 TB
▪ 3 QUAD CPU units → 300 cores

Data size

- Data are expected to be taken without zero skipping and exploiting loss-less compression and the system has been designed to support up to 100 Hz of beam triggers without zero-skipping and no compression

- 7680 channels, 10k samples in a drift windows of 4ms→ 146.8MB/events, No zero skipping

- Beam rate: 100Hz

- Data flow= 14.3 GB/s (without compression), 1.43 GB/s (with compression)

Huffman lossless compression can reduce the non-zero-skipped charge readout data volume by at least a factor 10 (S/N for double phase ~100:1, small noise fluctuations in absolute ADC counts)

- Light readout does not change in a significant way this picture (<0.5 GB/s)

→ Integrated internal local DAQ bandwidth on the "20 GB/s scale" in order to have a robust safety factor for concurrent read/write

Local data buffer ~ 1000TB (no zero skipping, no compression), also used for local processing

- 100 M triggers expected to be taken in 120 days of beam time in 2018

- If totally stored in non-zero-skipped, lossless compression format
  (assuming Huffman, factor 10 compression: 15MB/event) → 2.4 PB
  + cosmic runs and technical tests

- Requested link from online-storage to CERN computing division at 20 Gbps, compatible with 100 Hz non-zero-skipped, Huffman compressed (factor 10) data flow.

- This link would allow to transfer the entire beam triggers data volume with a typical occupancy of less or equal than 80%.

- The availability of a large local buffer allows as well to release the disk cashing requirements at the other end of the data link at the computing division being consistent with a dilution of the beam data transfer over the periods during which the experiment is not having beam time.

Online storage/processing farm motivation :

SPSC report, April 2016

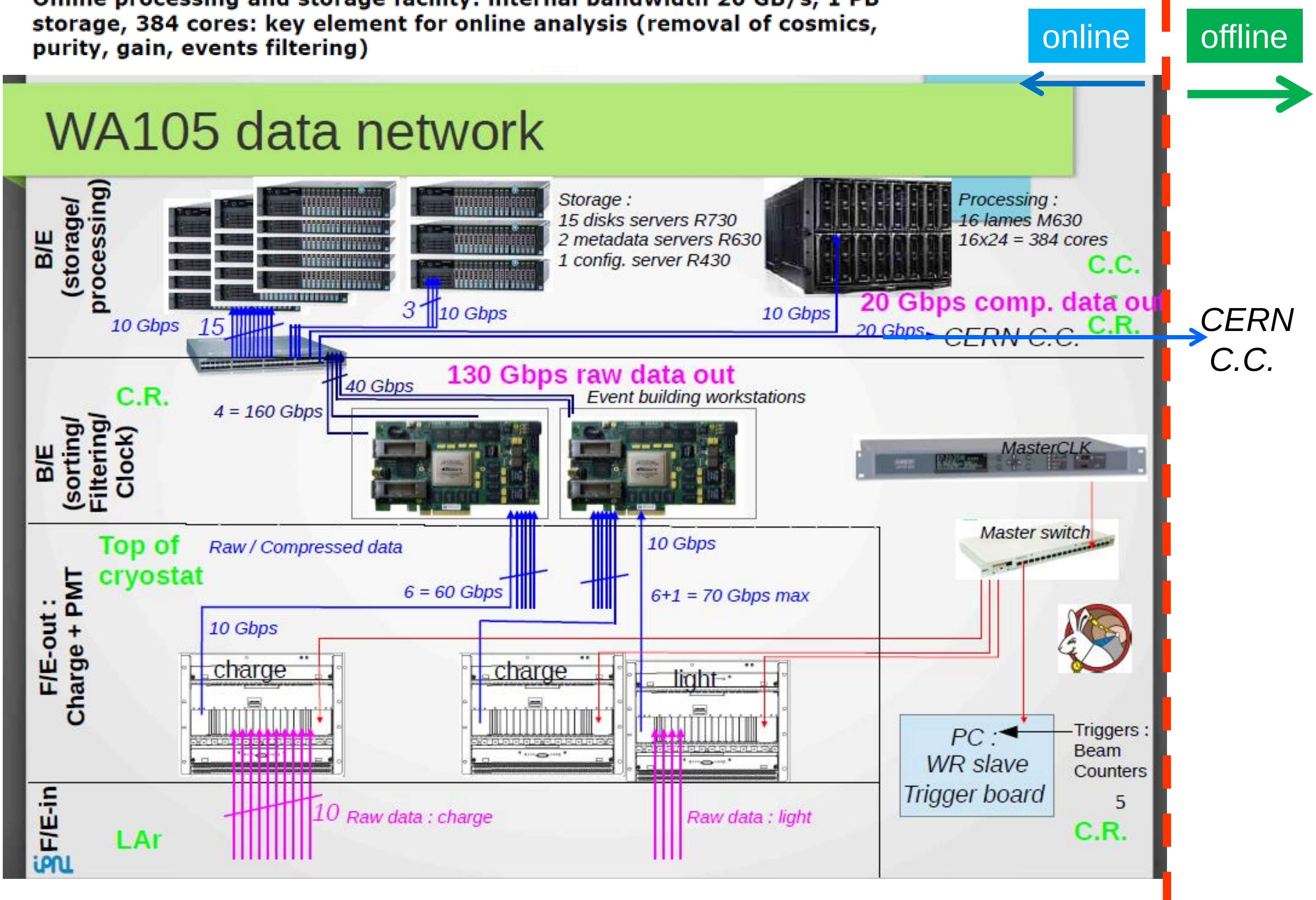
6x6x6

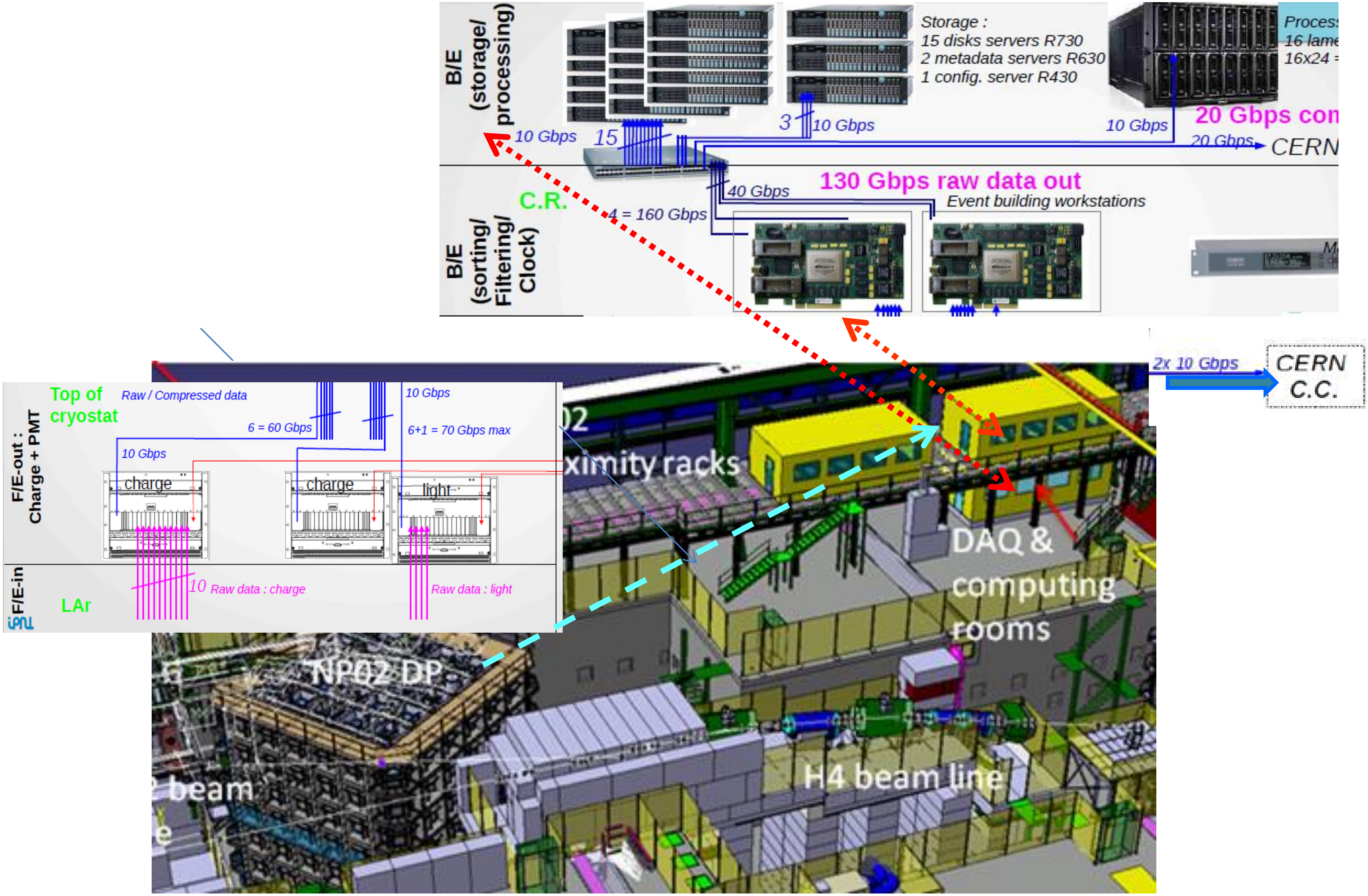
The local bandwidth of 20 GB/s also allows comfortable concurrent reading and writing access to the compressed data on the local storage system for online analysis. Data transfer to the IT division should happen by clustering the events in files having dimensions of a few gigabytes. This file size is needed for an efficient storage on the Castor system at the computing center. The online storage facility has also the task of buffering the events and formatting them for transfer on this typical file size.

In addition to the storage buffers requirement described above, the online storage processing farm allows for the following functionalities:

- Completion of event building by connecting the data flows of the two back-end systems
- Fast event reconstruction and disentangling of cosmic rays tracks segments by using also the LRO timing information
- Selection of a subsample of the cosmic ray tracks overlapped to beam events for online purity analysis and detector gain monitoring
- General online data quality checks
- Events filtering and formatting for final storage

Online processing and storage facility: internal bandwidth 20 GB/s, 1 PB storage, 384 cores: key element for online analysis (removal of cosmics, purity, gain, events filtering)



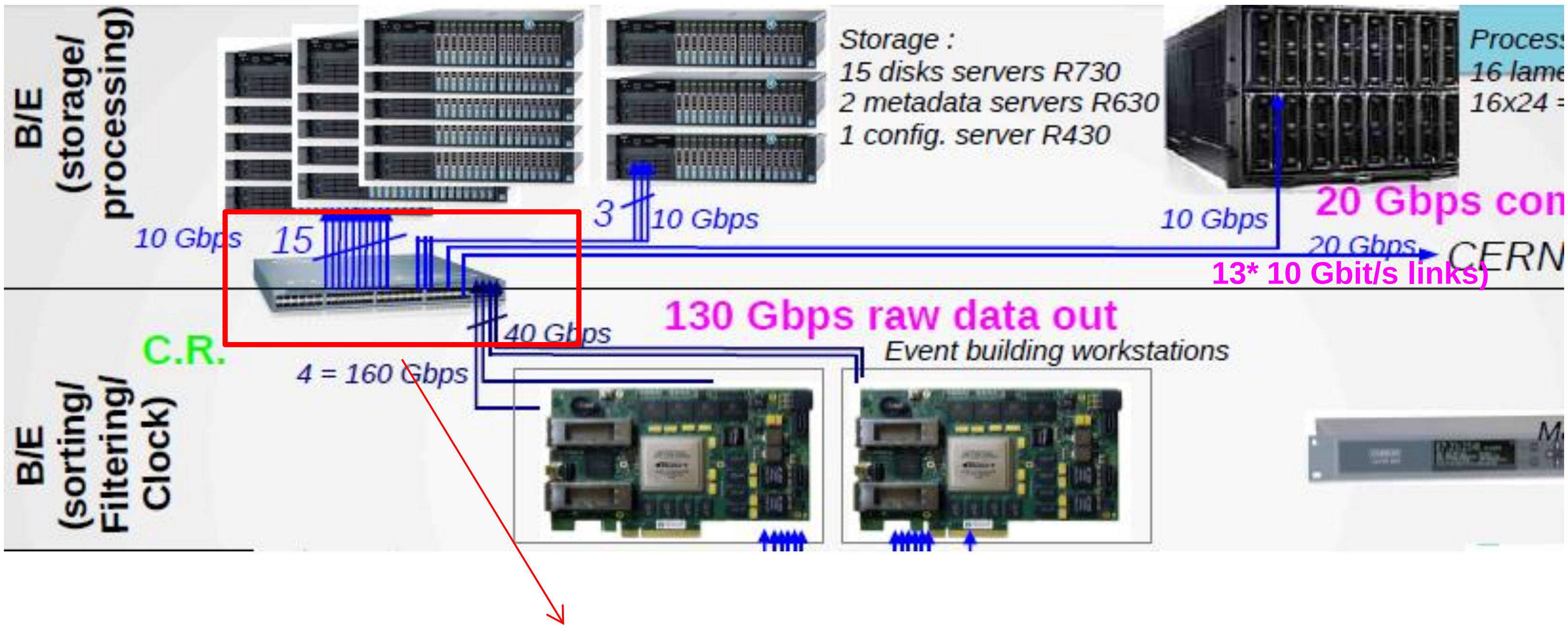

B/E (storage/ processing)
Storage :
15 disks servers R730
2 metadata servers R630
1 config. server R430
10 Gbps
15
3
10 Gbps
10 Gbps
20 Gbps
CERN
130 Gbps raw data out
Event building workstations
40 Gbps
4 = 160 Gbps
C.R.
B/E (sorting/ Filtering/ Clock)
2x 10 Gbps
CERN C.C.
F/E-out : Charge + PMT
Top of cryostat
Raw / Compressed data
6 = 60 Gbps
10 Gbps
6+1 = 70 Gbps max
10 Gbps
charge
charge
light
F/E-in
LAr
10
Raw data : charge
Raw data : light
DAQ &
computing
rooms
NP02 DP
H4 beam line

A data flow of 12 Gbps (compressed) has to be treated by the online storage farm

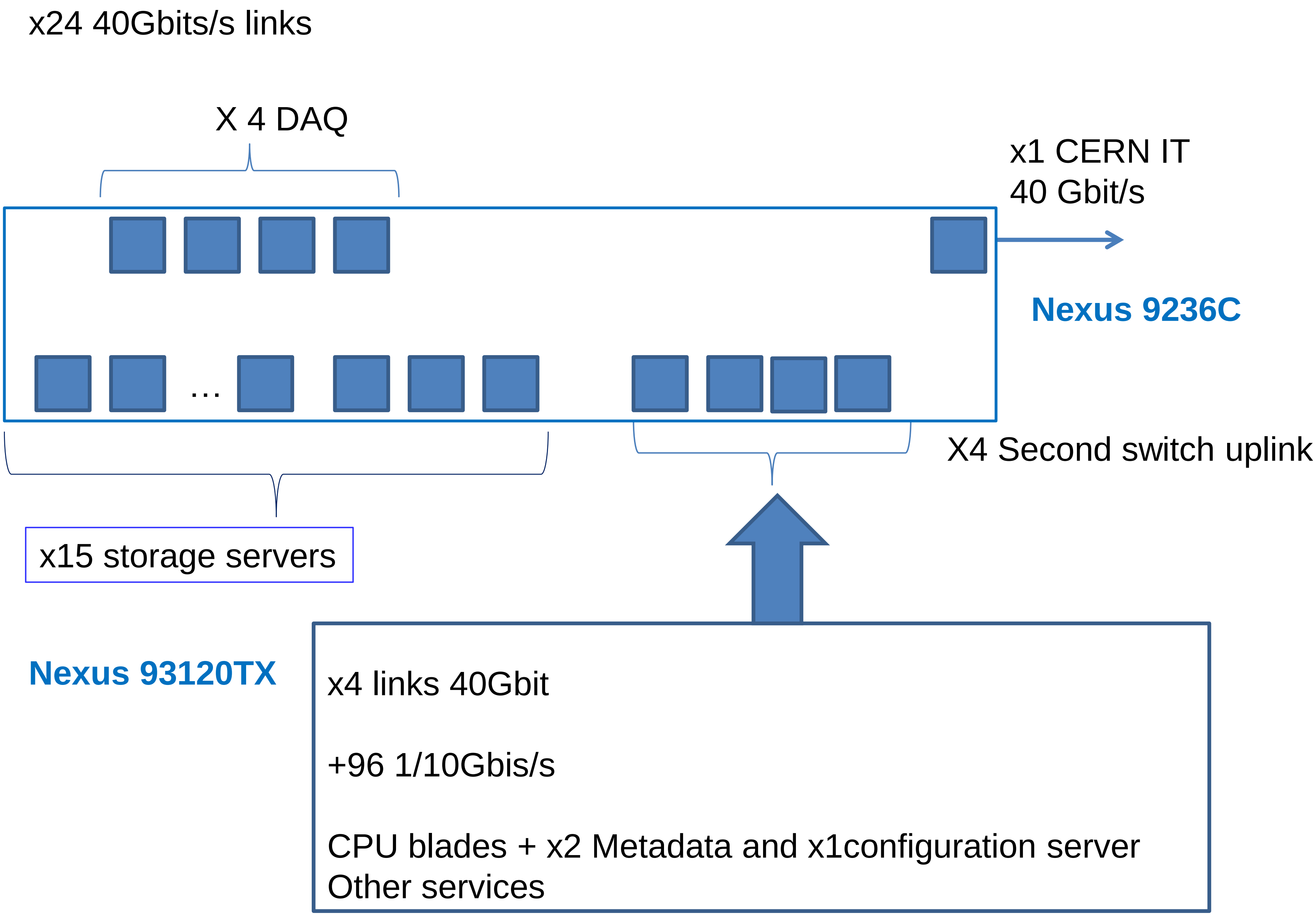
Data storage is distributed on 15 servers R730xd, each one including 16 disks of 6PB

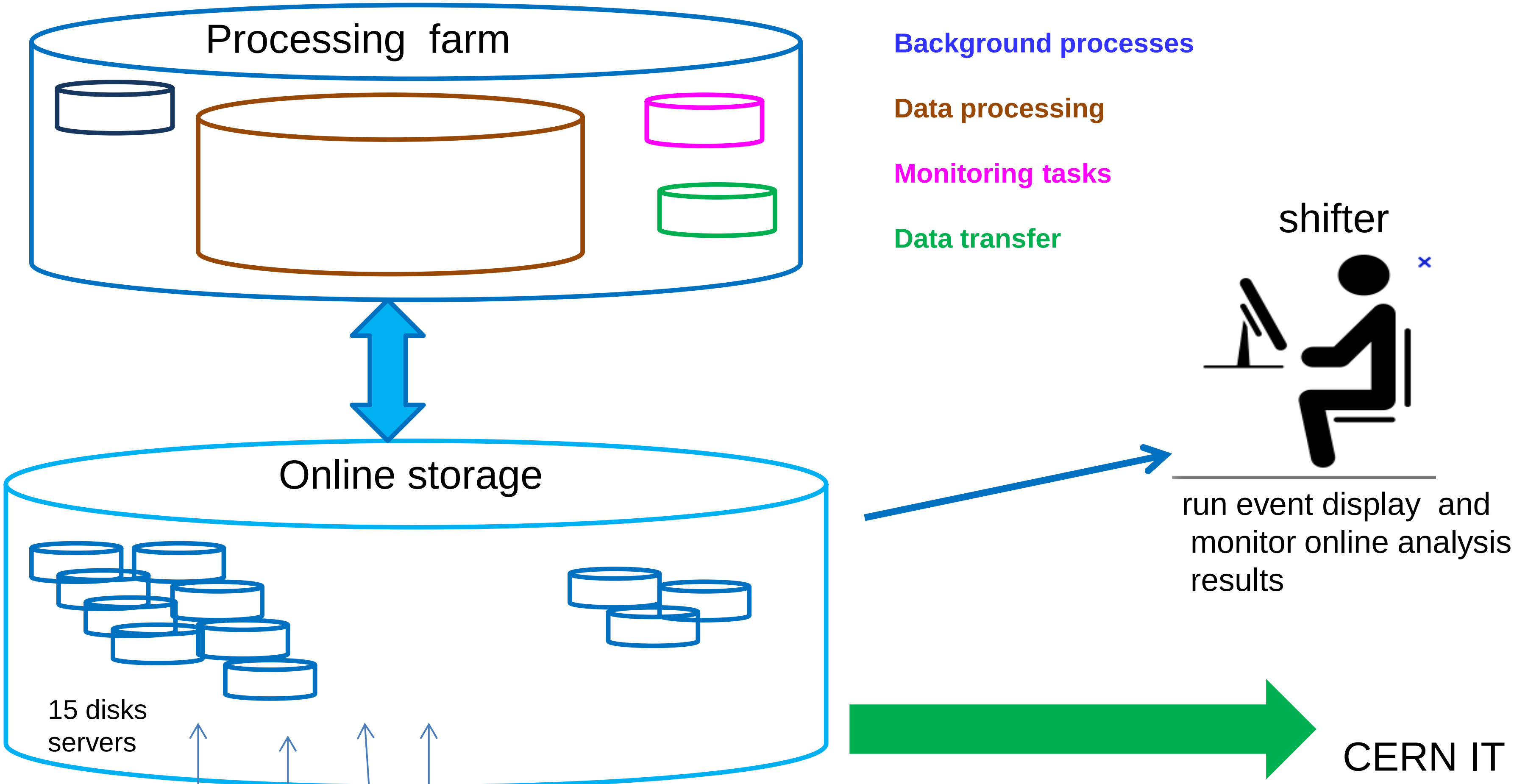
The system also includes 2 MetaData Servers (MDS, DELL R630) +1 configuration Server (DELL 430)

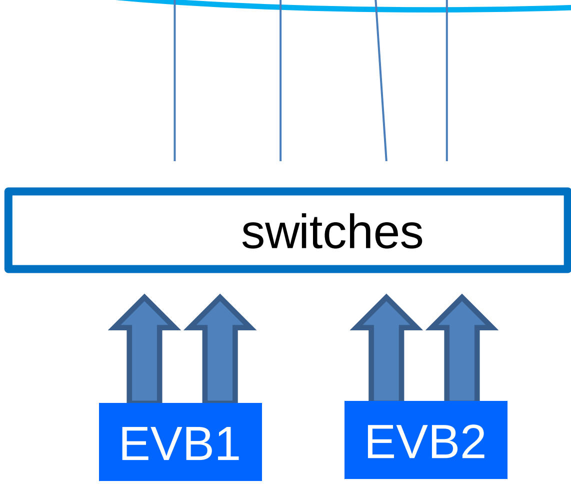
The local storage is based on EOS



see next page

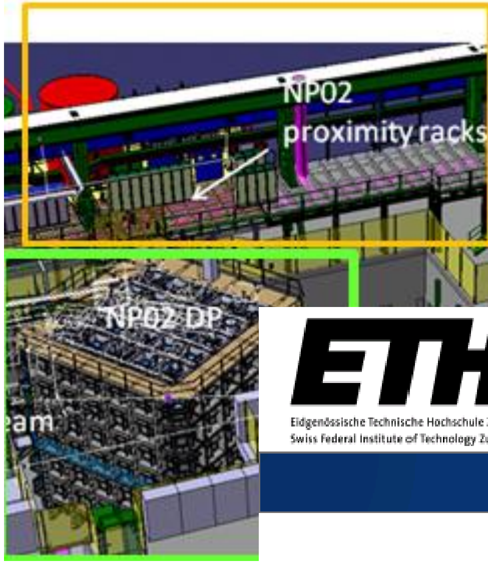

x24 40Gbits/s links
X 4 DAQ
x1 CERN IT
40 Gbit/s
Nexus 9236C
…
X4 Second switch uplink
x15 storage servers
Nexus 93120TX
x4 links 40Gbit
+96 1/10Gbis/s
CPU blades + x2 Metadata and x1configuration server
Other services





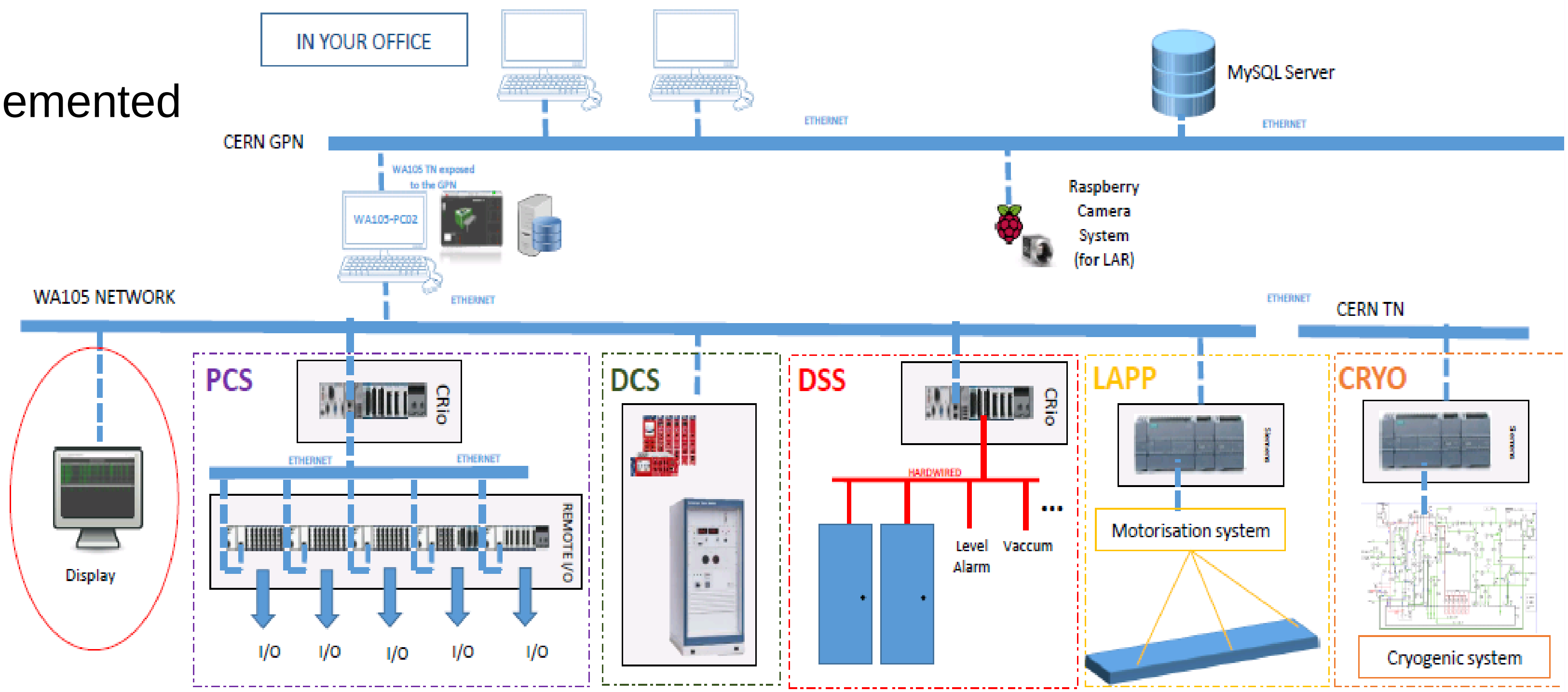
private network(s), with a local bandwidth of 20GB/s
to allow the transfer of compressed data to local storage system and concurrent reading and writing access for online analysis

CERN GENERAL PURPOSE NETWORK for access to machines ?



Slow Control

Scheme implemented on 3x1x1



Slow control database is accessible from